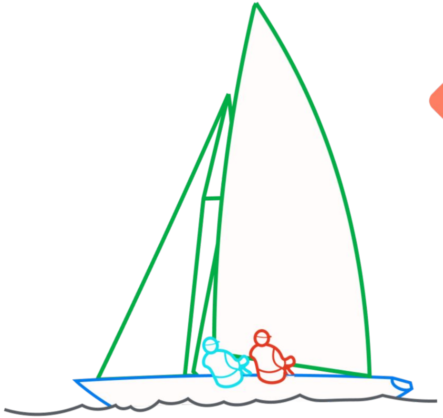
Airflow @Gojek
Streamlining Data Processing for Tableau Dashboards

Wanda Kinasih

BI @Gojek



Airflow Summit

Let's flow together

September 19-21, 2023,
Toronto, Canada

Hi! I’m Wanda!

Wanda Kinasih

- BI Analyst since 2016
- Now working as **BI Lead for Consumer Platform, Gojek**
- Experienced at:
  - SQL, Python
  - Data Visualisation using Tableau, Google Data Studio, Metabase
  - A/B Testing experiment
  - Google Cloud Project
  - Airflow, Pentaho
- Tableau Desktop Specialist Certified

Agenda

- Gojek Introduction
- Gojek Data Platform
- The Power of Airflow and Tableau Integration

gojek

A goto group operating company

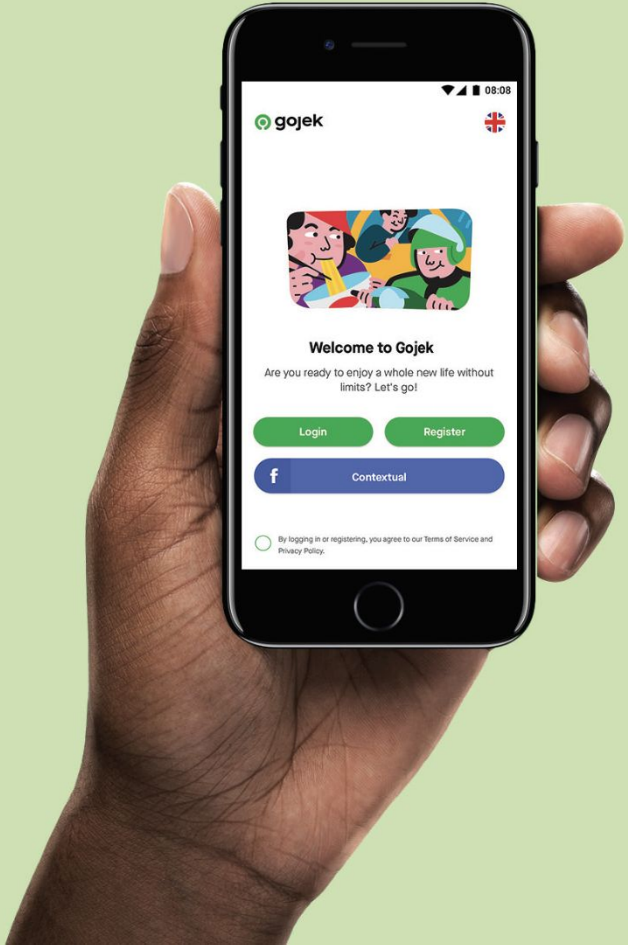
goto

VISION

Become the Micro-Entrepreneurs Hero Brand.

MISSION

Create & scale up positive socio-economic impact on the ecosystem of users, driver partners, business & SMEs, as-well-as service providers.



Welcome to Gojek

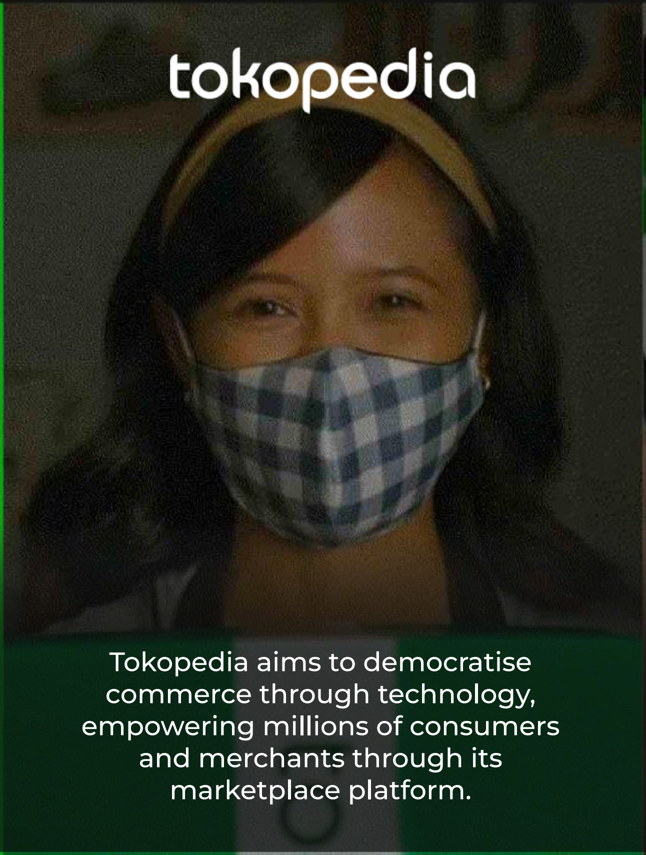
Are you ready to enjoy a whole new life without limits? Let's go!

○ By logging in or registering, you agree to our Terms of Service and Privacy Policy.

goto

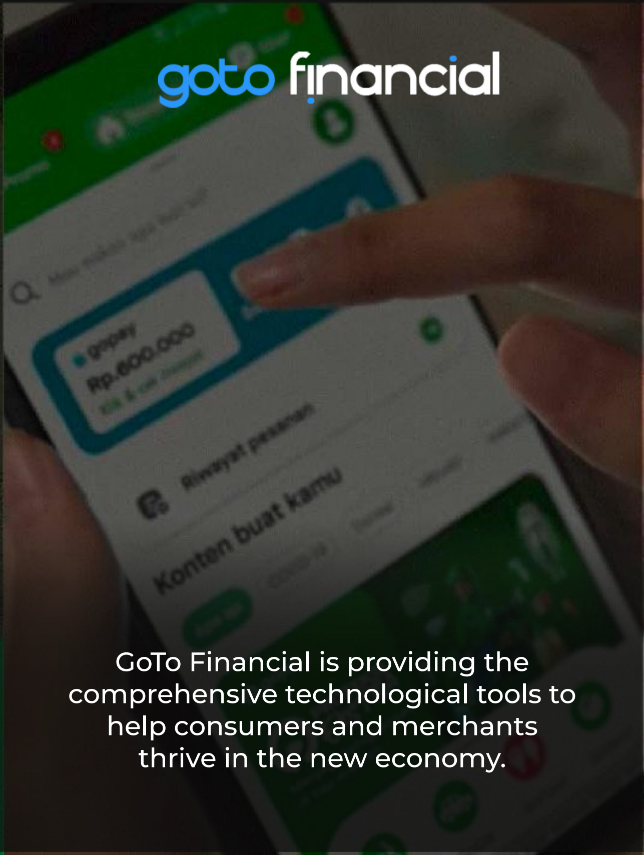
gojek

Gojek aims to empower micro-entrepreneurs to make cities more accessible and engaging.



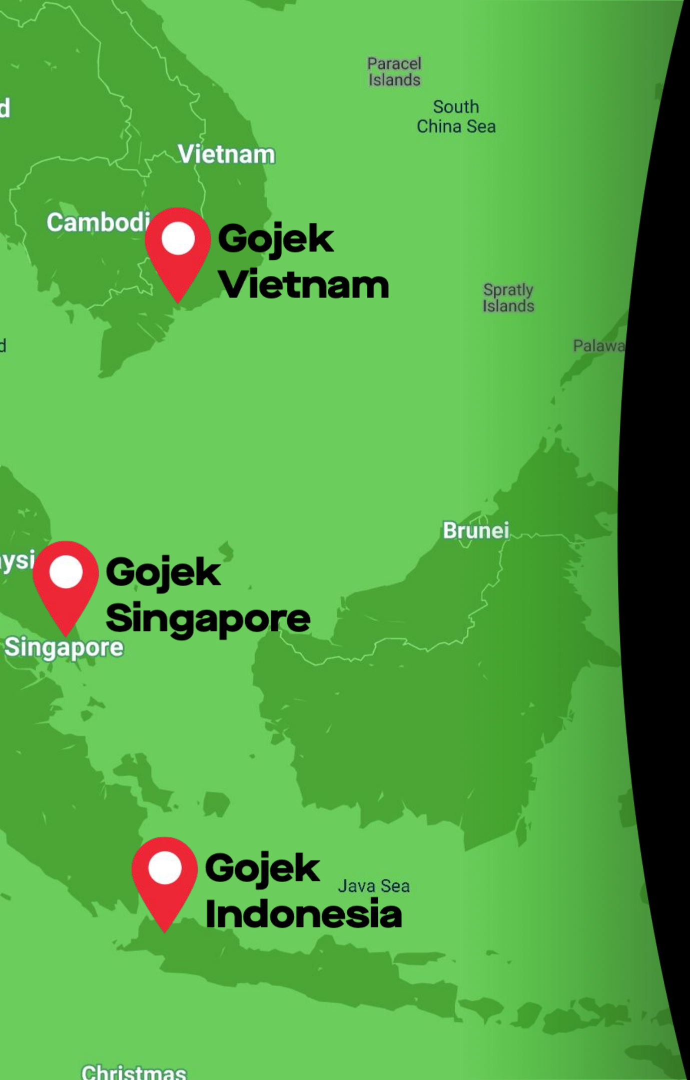
tokopedia

Tokopedia aims to democratise commerce through technology, empowering millions of consumers and merchants through its marketplace platform.



goto financial

GoTo Financial is providing the comprehensive technological tools to help consumers and merchants thrive in the new economy.

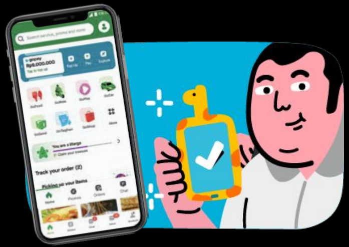


OUR FOOTPRINT IN SOUTHEAST ASIA

Founded in Indonesia, Gojek now operates in **three** Southeast Asian countries

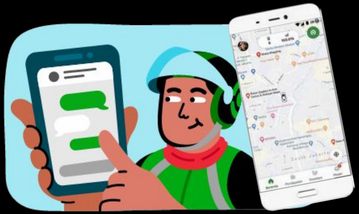
3 APPS:

Consumer, Merchant Partner & Driver Partner



Fulfils daily needs

Increases turnover & business scale



Optimizes the productivity of driver partners

THE JOURNEY SO FAR...

2010

Gojek started commercial operations.

2015

Launched on-demand services app in Indonesia

2016

Launched GoPay

2021

Entered Vietnam & Singapore

2021

United with Tokopedia to create GoTo

The “go to” ecosystem for daily life combining on-demand e-commerce & financial tech services

MOBILITY

GoCar
Car ride-hailing service.

GoRide
Motorcycle taxi (ojek) ride-hailing service.

GoCar Protect+
Extra protection to feel safe on a trip.

GoBluebird
Bluebird taxi booking service.

GoTransit
Multi-modal journey planner solution.

GoCorp
Platform for corporate clients to easily access and monitor business-related trips for their employees.

FOOD DELIVERY

GoMart
On-demand delivery from grocery and convenience stores.

GoFood
Food delivery service that provides consumers with convenient access to the best food options.

Cloud Kitchen
Shared kitchens for preparation of delivery-only meals.

LOGISTICS

GoSend
C2C product that provides consumers with fast and hassle-free instant and same-day delivery services.

GoSendAPI
A B2B2C delivery service offered specifically for business partners.

GoBox
On-demand truck logistics service for large-sized deliveries.

GoShop
On-demand personal concierge service allowing consumers to shop for items and have them delivered within hours.

OUR IMPACT

Economy

Gojek contributed **IDR 249 T** to the national economy (equivalent to **1.6% of Indonesia's GDP** in 2020)

>1 million
GoFood merchant partners

>2.6 million
driver partners

Driver Partners

Gojek driver partners remain resilient during the pandemic.

Driver partners have experiences significant recovery through an increase in income of:

24% For GoRide partners

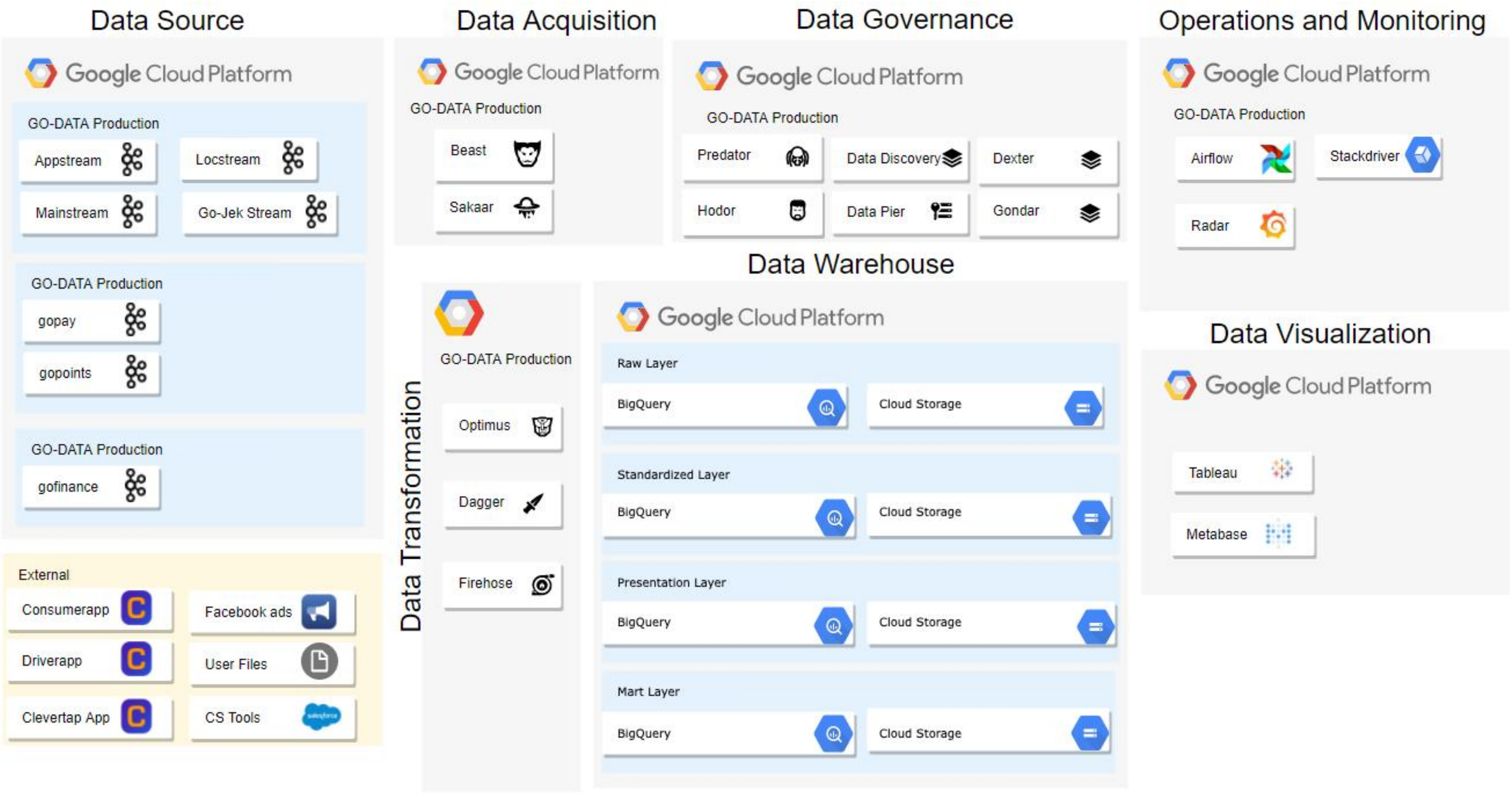
18% For GoRide partners

Compared to the beginning of the pandemic

4 out of 5 partners still have an income to support themselves and their families

2 out of 3 partners feel the benefit from the time flexibility in their partnership with Gojek.

Gojek Data Platform



Data Source

Google Cloud Platform

GO-DATA Production

- Appstream
- Locstream
- Mainstream
- Go-Jek Stream

GO-DATA Production

- gopay
- gopoints

GO-DATA Production

- gofinance

External

- Consumerapp
- Facebook ads
- Driverapp
- User Files
- Clevertap App
- CS Tools

Data Acquisition

Google Cloud Platform

GO-DATA Production

- Beast
- Sakaar

Data Governance

Google Cloud Platform

GO-DATA Production

- Predator
- Data Discovery
- Dexter
- Hodor
- Data Pier
- Gondar

Operations and Monitoring

Google Cloud Platform

GO-DATA Production

- Airflow
- Stackdriver
- Radar

Data Transformation

GO-DATA Production

- Optimus
- Dagger
- Firehose

Data Warehouse

Google Cloud Platform

Raw Layer

- BigQuery
- Cloud Storage

Standardized Layer

- BigQuery
- Cloud Storage

Presentation Layer

- BigQuery
- Cloud Storage

Mart Layer

- BigQuery
- Cloud Storage

Data Visualization

Google Cloud Platform

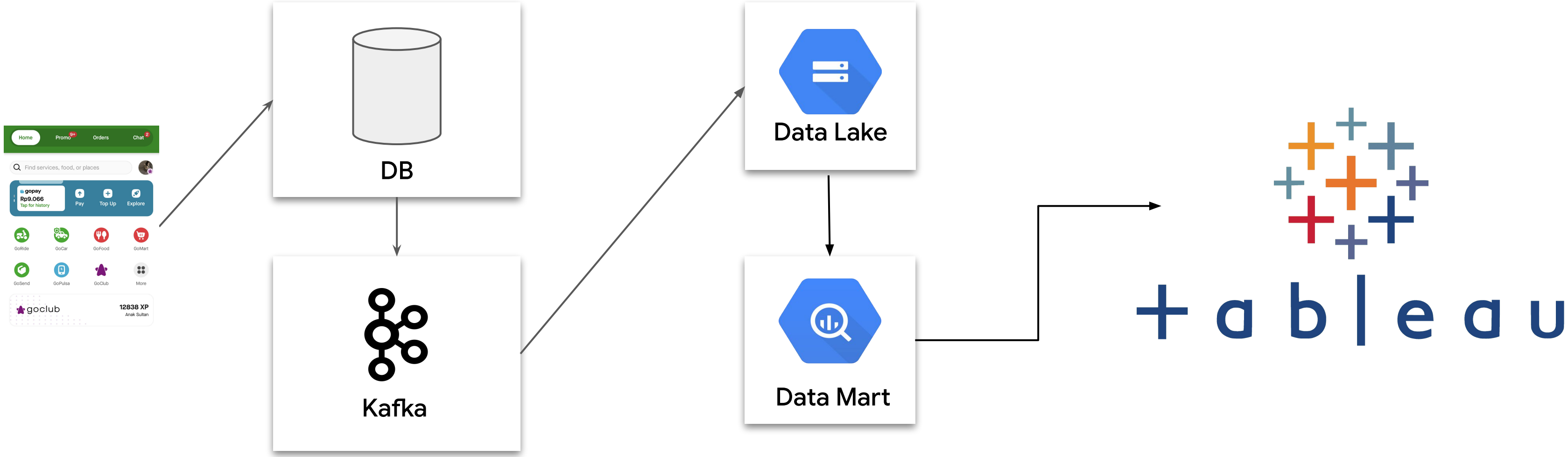
- Tableau
- Metabase

(Simplified) Data Pipeline

Source
(product)

Store
(DWH)

Analyse
(Analyst, Business)



DB

Kafka

Data Lake

Data Mart

Importance of Efficient Data Processing and Visualization

1. **Informed Decision Making**

   Efficient data processing and visualization enable organizations to quickly turn raw data into meaningful insights.

2. **Faster Problem Solving**

   By analyzing data in real-time and visualizing it in a comprehensible manner, organizations can identify issues early, troubleshoot efficiently, and minimize downtime.

3. **Scalability and Performance**

   Properly processed and visualized data allows systems to handle larger datasets without compromising speed or accuracy.

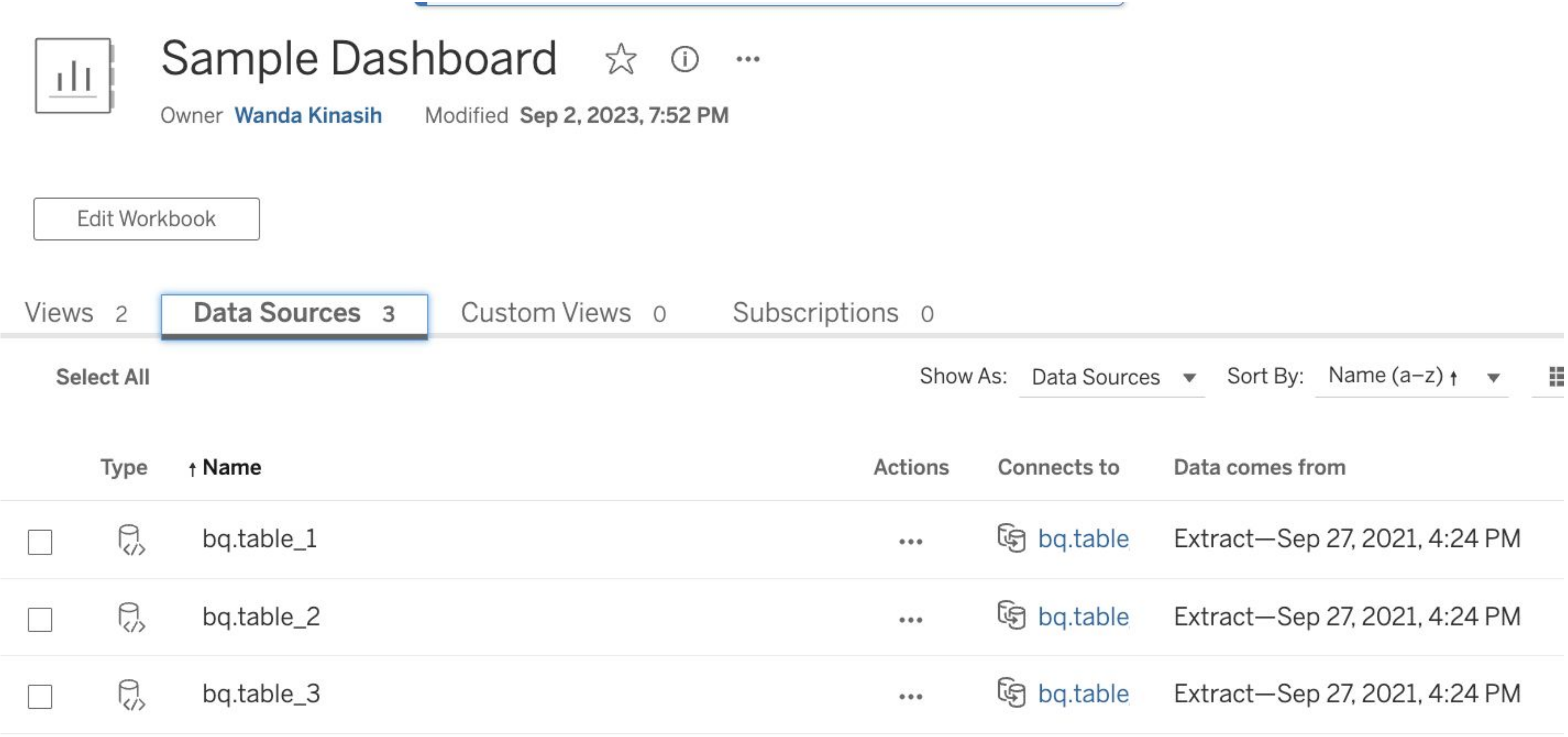
4. **Data Quality Assurance**

   Instantly detect inconsistencies, errors, and outliers, allowing data engineers to maintain high-quality datasets.

Gojek Tableau Dashboards

- ✓ >300 Data Sources
- ✓ >800 Dependencies
- ✓ >400 Daily Views
- ✓ >500 Active Users
- ✓ 55 Dashboard Creators

Lots of Data Sources in each Dashboard

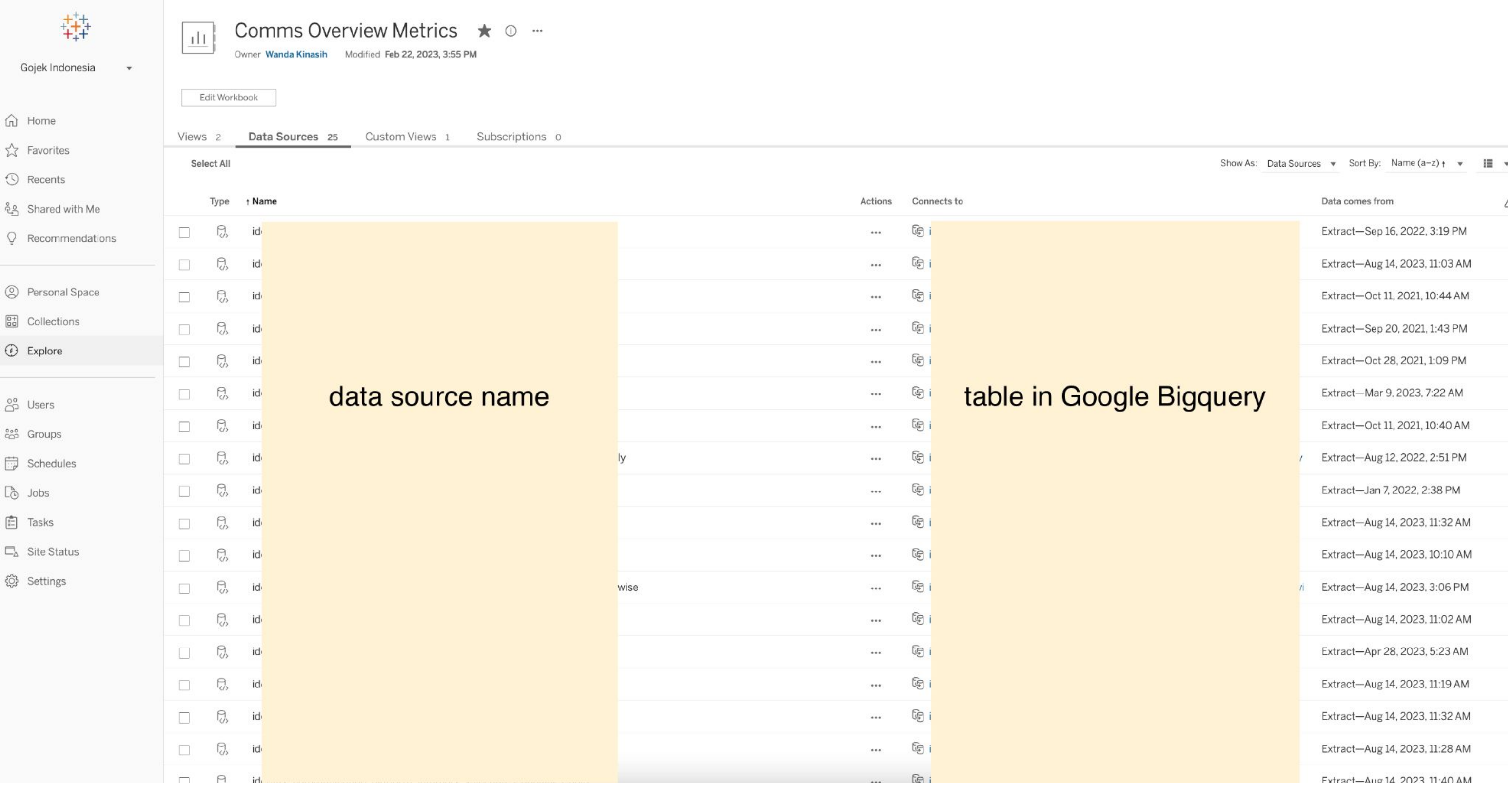


Owner **Wanda Kinasih** Modified **Sep 2, 2023, 7:52 PM**

Show As: Data Sources ▾ Sort By: Name (a–z) ↑ ▾

	Type	↑ Name	Actions	Connects to	Data comes from
□		bq.table_1	•••	bq.table	Extract—Sep 27, 2021, 4:24 PM
□		bq.table_2	•••	bq.table	Extract—Sep 27, 2021, 4:24 PM
□		bq.table_3	•••	bq.table	Extract—Sep 27, 2021, 4:24 PM

Huge Data Sources in each Dashboard



Gojek Indonesia

- Home
- Favorites
- Recents
- Shared with Me
- Recommendations
- Personal Space
- Collections
- Explore
- Users
- Groups
- Schedules
- Jobs
- Tasks
- Site Status
- Settings

Comms Overview Metrics

Owner Wanda Kinasih Modified Feb 22, 2023, 3:55 PM

Edit Workbook

Views 2 | Data Sources 25 | Custom Views 1 | Subscriptions 0

Select All

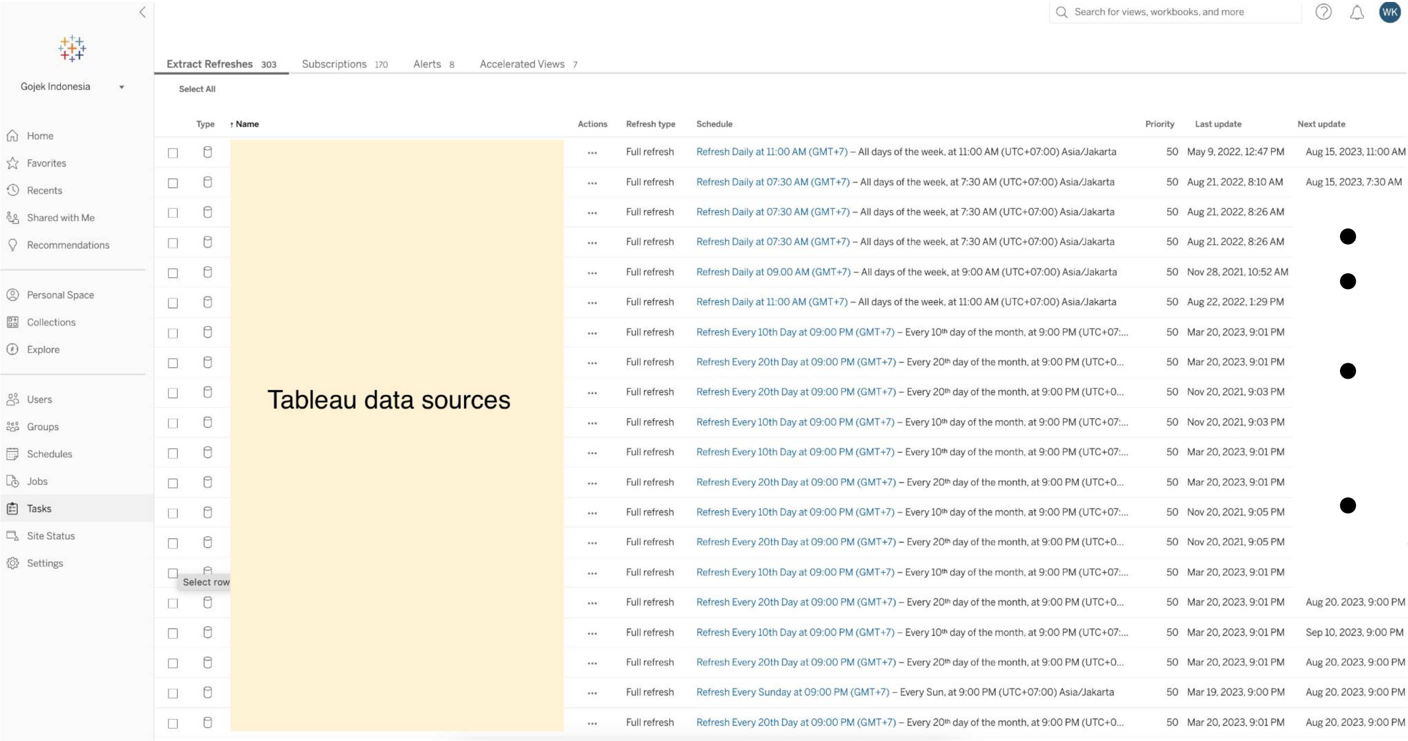
Show As: Data Sources Sort By: Name (a–z) ↑

	Type	↑ Name	Actions	Connects to	Data comes from
□		id	···	i	Extract—Sep 16, 2022, 3:19 PM
□		id	···	i	Extract—Aug 14, 2023, 11:03 AM
□		id	···	i	Extract—Oct 11, 2021, 10:44 AM
□		id	···	i	Extract—Sep 20, 2021, 1:43 PM
□		id	···	i	Extract—Oct 28, 2021, 1:09 PM
□		id	···	i	Extract—Mar 9, 2023, 7:22 AM
□		id	···	i	Extract—Oct 11, 2021, 10:40 AM
□		id ly	···	i	Extract—Aug 12, 2022, 2:51 PM
□		id	···	i	Extract—Jan 7, 2022, 2:38 PM
□		id	···	i	Extract—Aug 14, 2023, 11:32 AM
□		id	···	i	Extract—Aug 14, 2023, 10:10 AM
□		id wise	···	i	Extract—Aug 14, 2023, 3:06 PM
□		id	···	i	Extract—Aug 14, 2023, 11:02 AM
□		id	···	i	Extract—Apr 28, 2023, 5:23 AM
□		id	···	i	Extract—Aug 14, 2023, 11:19 AM
□		id	···	i	Extract—Aug 14, 2023, 11:32 AM
□		id	···	i	Extract—Aug 14, 2023, 11:28 AM
□		id	···	i	Extract—Aug 14, 2023, 11:40 AM

data source name

table in Google Bigquery

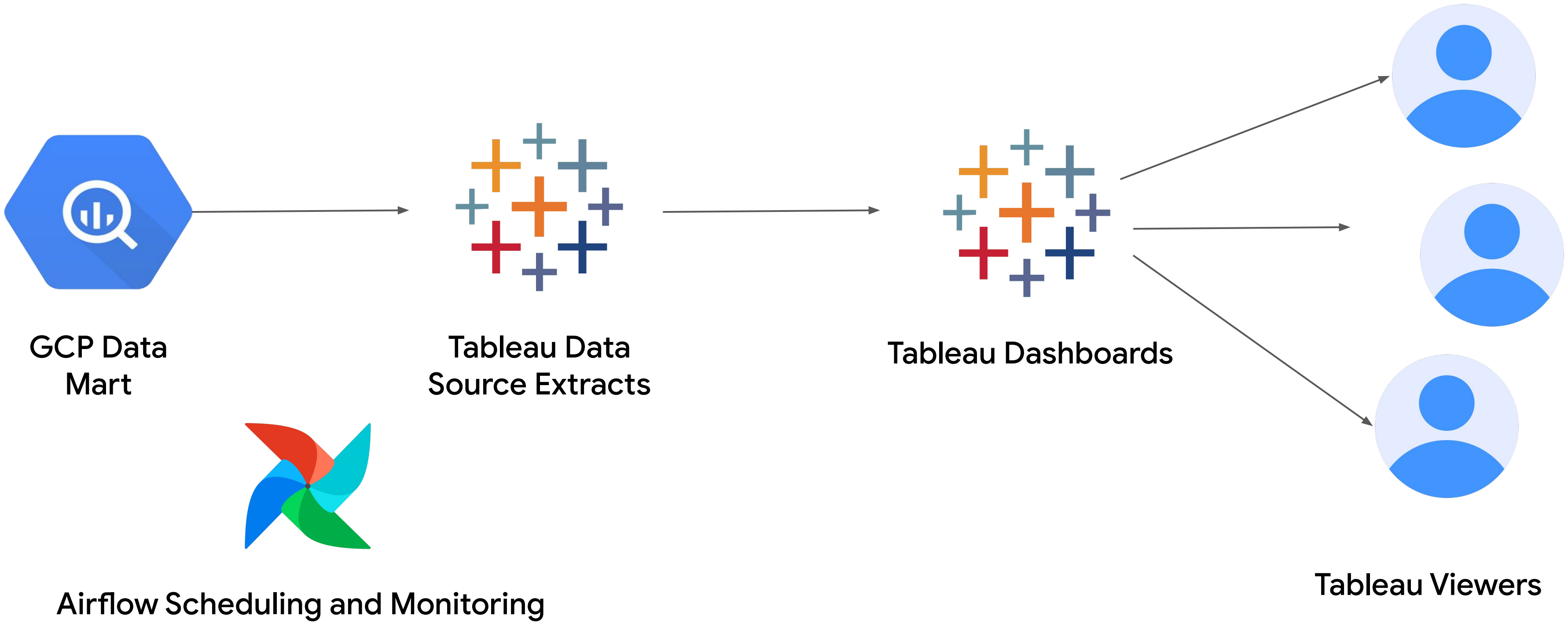
Tableau Built-in Scheduler



	Type	↑ Name	Actions	Refresh type	Schedule	Priority	Last update	Next update
□			...	Full refresh	Refresh Daily at 11:00 AM (GMT+7) – All days of the week, at 11:00 AM (UTC+07:00) Asia/Jakarta	50	May 9, 2022, 12:47 PM	Aug 15, 2023, 11:00 AM
□			...	Full refresh	Refresh Daily at 07:30 AM (GMT+7) – All days of the week, at 7:30 AM (UTC+07:00) Asia/Jakarta	50	Aug 21, 2022, 8:10 AM	Aug 15, 2023, 7:30 AM
□			...	Full refresh	Refresh Daily at 07:30 AM (GMT+7) – All days of the week, at 7:30 AM (UTC+07:00) Asia/Jakarta	50	Aug 21, 2022, 8:26 AM	
□			...	Full refresh	Refresh Daily at 07:30 AM (GMT+7) – All days of the week, at 7:30 AM (UTC+07:00) Asia/Jakarta	50	Aug 21, 2022, 8:26 AM	
□			...	Full refresh	Refresh Daily at 09.00 AM (GMT+7) – All days of the week, at 9:00 AM (UTC+07:00) Asia/Jakarta	50	Nov 28, 2021, 10:52 AM	
□			...	Full refresh	Refresh Daily at 11:00 AM (GMT+7) – All days of the week, at 11:00 AM (UTC+07:00) Asia/Jakarta	50	Aug 22, 2022, 1:29 PM	
□			...	Full refresh	Refresh Every 10th Day at 09:00 PM (GMT+7) – Every 10th day of the month, at 9:00 PM (UTC+07:...	50	Mar 20, 2023, 9:01 PM	
□			...	Full refresh	Refresh Every 20th Day at 09:00 PM (GMT+7) – Every 20th day of the month, at 9:00 PM (UTC+0...	50	Mar 20, 2023, 9:01 PM	
□			...	Full refresh	Refresh Every 20th Day at 09:00 PM (GMT+7) – Every 20th day of the month, at 9:00 PM (UTC+0...	50	Nov 20, 2021, 9:03 PM	
□			...	Full refresh	Refresh Every 10th Day at 09:00 PM (GMT+7) – Every 10th day of the month, at 9:00 PM (UTC+07:...	50	Nov 20, 2021, 9:03 PM	
□			...	Full refresh	Refresh Every 10th Day at 09:00 PM (GMT+7) – Every 10th day of the month, at 9:00 PM (UTC+07:...	50	Mar 20, 2023, 9:01 PM	
□			...	Full refresh	Refresh Every 20th Day at 09:00 PM (GMT+7) – Every 20th day of the month, at 9:00 PM (UTC+0...	50	Mar 20, 2023, 9:01 PM	
□			...	Full refresh	Refresh Every 10th Day at 09:00 PM (GMT+7) – Every 10th day of the month, at 9:00 PM (UTC+07:...	50	Nov 20, 2021, 9:05 PM	
□			...	Full refresh	Refresh Every 20th Day at 09:00 PM (GMT+7) – Every 20th day of the month, at 9:00 PM (UTC+0...	50	Nov 20, 2021, 9:05 PM	
□			...	Full refresh	Refresh Every 10th Day at 09:00 PM (GMT+7) – Every 10th day of the month, at 9:00 PM (UTC+07:...	50	Mar 20, 2023, 9:01 PM	
□			...	Full refresh	Refresh Every 20th Day at 09:00 PM (GMT+7) – Every 20th day of the month, at 9:00 PM (UTC+0...	50	Mar 20, 2023, 9:01 PM	Aug 20, 2023, 9:00 PM
□			...	Full refresh	Refresh Every 10th Day at 09:00 PM (GMT+7) – Every 10th day of the month, at 9:00 PM (UTC+07:...	50	Mar 20, 2023, 9:01 PM	Sep 10, 2023, 9:00 PM
□			...	Full refresh	Refresh Every 20th Day at 09:00 PM (GMT+7) – Every 20th day of the month, at 9:00 PM (UTC+0...	50	Mar 20, 2023, 9:01 PM	Aug 20, 2023, 9:00 PM
□			...	Full refresh	Refresh Every Sunday at 09:00 PM (GMT+7) – Every Sun, at 9:00 PM (UTC+07:00) Asia/Jakarta	50	Mar 19, 2023, 9:00 PM	Aug 20, 2023, 9:00 PM
□			...	Full refresh	Refresh Every 20th Day at 09:00 PM (GMT+7) – Every 20th day of the month, at 9:00 PM (UTC+0...	50	Mar 20, 2023, 9:01 PM	Aug 20, 2023, 9:00 PM

- Only list of schedules
- Can’t set dependencies
- Can’t monitor each data source refresh easily
- Can’t see which job is failing

Integrating Tableau and Airflow



GCP Data Mart

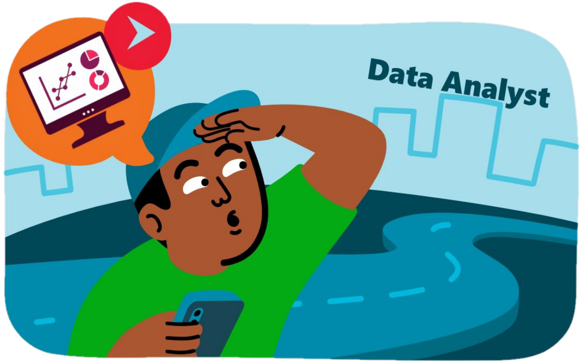
Tableau Data Source Extracts

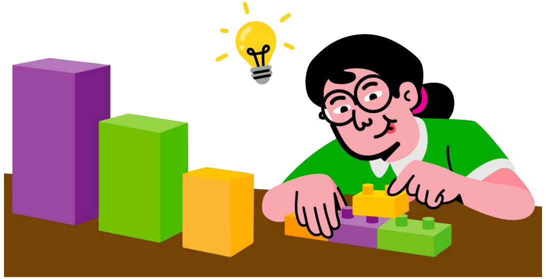
Tableau Dashboards

Tableau Viewers

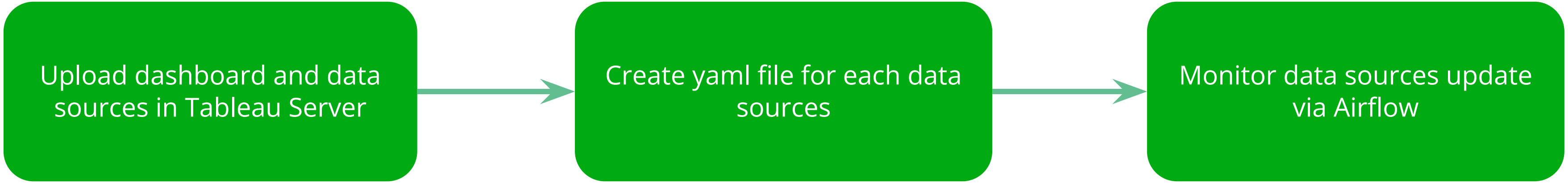
Airflow Scheduling and Monitoring

How To Make Sure Each Data Source Wait For Their Dependencies?





Set Up Schedule Easily via Airflow



Upload dashboard and data sources in Tableau Server

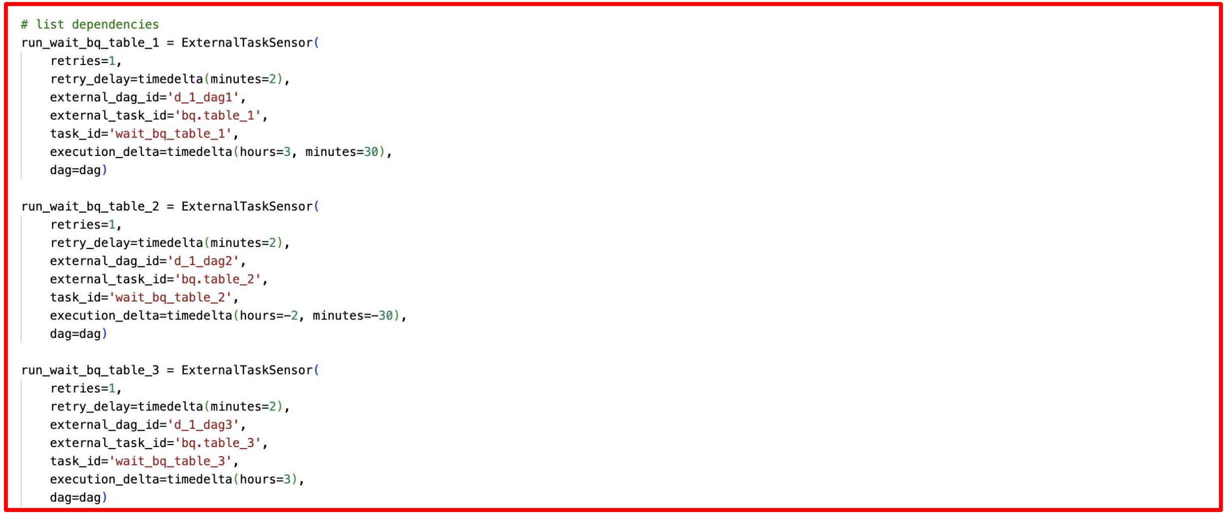
Create yaml file for each data sources

Monitor data sources update via Airflow

tableau

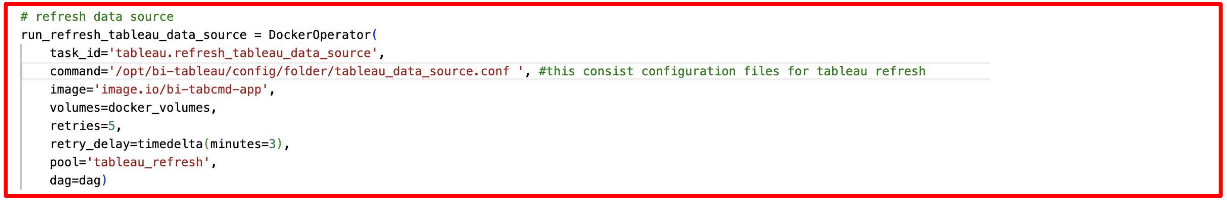
Apache Airflow

DAG Configuration (Py File)



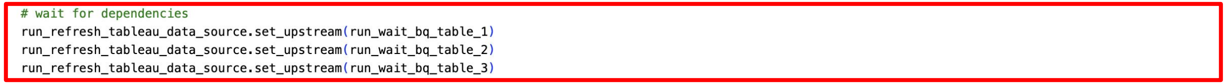
```
# list dependencies
run_wait_bq_table_1 = ExternalTaskSensor(
    retries=1,
    retry_delay=timedelta(minutes=2),
    external_dag_id='d_1_dag1',
    external_task_id='bq.table_1',
    task_id='wait_bq_table_1',
    execution_delta=timedelta(hours=3, minutes=30),
    dag=dag)

run_wait_bq_table_2 = ExternalTaskSensor(
    retries=1,
    retry_delay=timedelta(minutes=2),
    external_dag_id='d_1_dag2',
    external_task_id='bq.table_2',
    task_id='wait_bq_table_2',
    execution_delta=timedelta(hours=-2, minutes=-30),
    dag=dag)

run_wait_bq_table_3 = ExternalTaskSensor(
    retries=1,
    retry_delay=timedelta(minutes=2),
    external_dag_id='d_1_dag3',
    external_task_id='bq.table_3',
    task_id='wait_bq_table_3',
    execution_delta=timedelta(hours=3),
    dag=dag)
```

Dependencies in Google Cloud Bigquery



```
# refresh data source
run_refresh_tableau_data_source = DockerOperator(
    task_id='tableau.refresh_tableau_data_source',
    command='/opt/bi-tableau/config/folder/tableau_data_source.conf ', #this consist configuration files for tableau refresh
    image='image.io/bi-tabcmd-app',
    volumes=docker_volumes,
    retries=5,
    retry_delay=timedelta(minutes=3),
    pool='tableau_refresh',
    dag=dag)
```

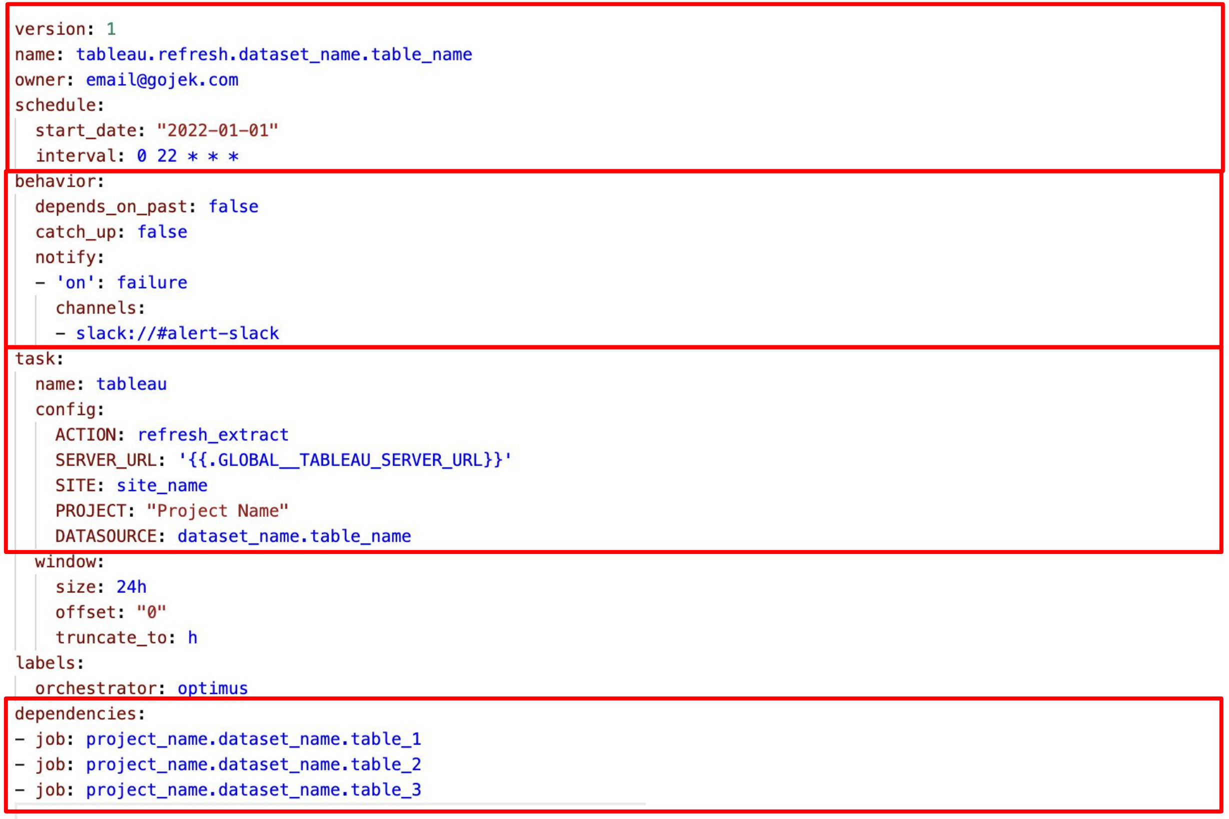
Tableau data source



```
# wait for dependencies
run_refresh_tableau_data_source.set_upstream(run_wait_bq_table_1)
run_refresh_tableau_data_source.set_upstream(run_wait_bq_table_2)
run_refresh_tableau_data_source.set_upstream(run_wait_bq_table_3)
```

Set upstream for each dependency

Simple Yaml File for Analysts and Business Users



```
version: 1
name: tableau.refresh.dataset_name.table_name
owner: email@gojek.com
schedule:
  start_date: "2022-01-01"
  interval: 0 22 * * *
behavior:
  depends_on_past: false
  catch_up: false
  notify:
  - 'on': failure
    channels:
    - slack://#alert-slack
task:
  name: tableau
  config:
    ACTION: refresh_extract
    SERVER_URL: '{{.GLOBAL__TABLEAU_SERVER_URL}}'
    SITE: site_name
    PROJECT: "Project Name"
    DATASOURCE: dataset_name.table_name
  window:
    size: 24h
    offset: "0"
    truncate_to: h
labels:
  orchestrator: optimus
dependencies:
- job: project_name.dataset_name.table_1
- job: project_name.dataset_name.table_2
- job: project_name.dataset_name.table_3
```

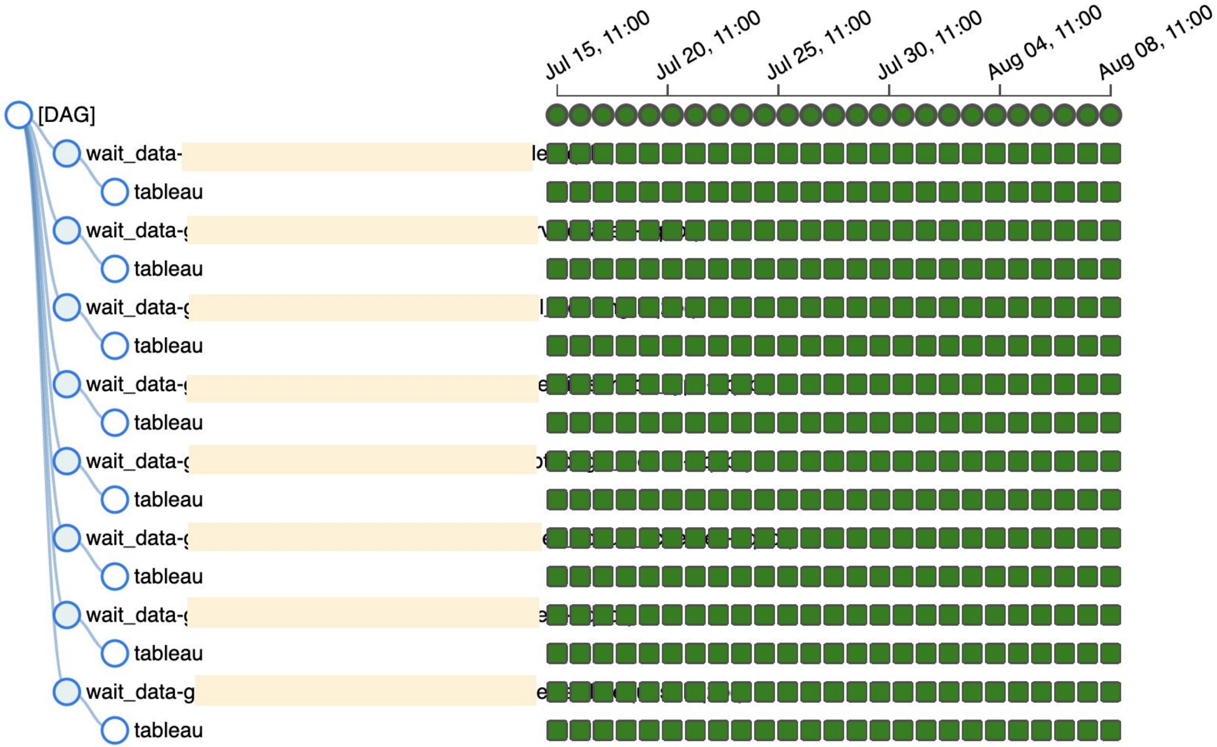
Dag name and schedule

Alerts

Tableau data source

Dependencies in Google Cloud Bigquery

Monitor Data Sources Easily via Airflow



Questions?

Let’s connect
https://www.linkedin.com/in/wandakinasih/